

Title:

Subtitle

©Patrick Hall 20XX

May 21, 2024

Abstract

1 Introduction

The National Institute of Standards and Technology Artificial Intelligence (AI) Risk Management Framework (RMF).[?]

2 Generative AI Governance

3 Generative AI Inventories

4 Generative AI Risk Tiers

5 Generative AI Risk Measurement

6 Generative AI Risk Management

Conclusion

Acknowledgments

Thank you to Bernie Siskin and Nick Schmidt of BLDS and Eric Sublett of Relman Colfax for formative discussions relating to GAI risk tiering.

Abbreviations

- AI: Artificial Intelligence
- AI RMF: Artificial Intelligence Risk Management Framework
- GAI: Generative AI
- RMF: Risk Management Framework

Appendix A: Example Generative AI–Trustworthy Characteristic Crosswalk

6.1 A.1: Trustworthy Characteristic to Generative AI Risk Crosswalk

Table 1: Trustworthy Characteristic to Generative AI Risk Crosswalk.

Accountable and Transparent	Explainable and Interpretable	Fair with Harmful Bias Managed	Privacy Enhanced
Data Privacy Environmental Human-AI Configuration Information Integrity Intellectual Property Value Chain and Component Integration	Human-AI Configuration Value Chain and Component Integration	Confabulation Environmental Human-AI Configuration Intellectual Property Obscene, Degrading, and/or Abusive Content Toxicity, Bias, and Homogenization Value Chain and Component Integration	Data Privacy Human-AI Configuration Information Security Intellectual Property Value Chain and Component Integration

Safe	Secure and Resilient	Valid and Reliable
CBRN Information Confabulation Dangerous or Violent Recommendations Data Privacy Environmental Human-AI Configuration Information Integrity Information Security Obscene, Degrading, and/or Abusive Content Value Chain and Component Integration	Dangerous or Violent Recommendations Data Privacy Human-AI Configuration Information Security Value Chain and Component Integration	Confabulation Human-AI Configuration Information Integrity Information Security Toxicity, Bias, and Homogenization Value Chain and Component Integration

6.2 A.2: Generative AI Risk to Trustworthy Characteristic Crosswalk

Table 2: Generative AI Risk to Trustworthy Characteristic Crosswalk.

CBRN Information	Confabulation	Dangerous or Violent Recommendations	Data Privacy
Safe	Fair with Harmful Bias Managed Safe Valid and Reliable	Safe Secure and Resilient	Accountable and Transparent Privacy Enhanced Safe Secure and Resilient
Environmental	Human-AI Configuration	Information Integrity	Information Security
Accountable and Transparent Fair with Harmful Bias Managed Safe	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable	Accountable and Transparent Safe Valid and Reliable	Privacy Enhanced Safe Secure and Resilient Valid and Reliable
Intellectual Property	Obscene, Degrading, and/or Abusive Content	Toxicity, Bias, and Homogenization	Value Chain and Component Integration
Accountable and Transparent Fair with Harmful Bias Managed Privacy Enhanced	Fair with Harmful Bias Managed Safe	Fair with Harmful Bias Managed Valid and Reliable	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Appendix B: Example Risk Tiers for Generative AI

Appendix C: List of Publicly Available Model Testing Suites (“Evals”)

C.1: Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic

Table 3: Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic.

Accountable and Transparent
An Evaluation on Large Language Model Outputs: Discourse and Memorization (see Appendix B) Big-bench: Truthfulness DecodingTrust: Machine Ethics Evaluation Harness: ETHICS HELM: Copyright Mark My Words
Fair with Harmful Bias Managed
BELEBELE Big-bench: Low-resource language, Non-English, Translation Big-bench: Social bias, Racial bias, Gender bias, Religious bias Big-bench: Toxicity DecodingTrust: Fairness DecodingTrust: Stereotype Bias DecodingTrust: Toxicity C-Eval (Chinese evaluation suite) Evaluation Harness: CrowS-Pairs Evaluation Harness: ToxiGen Finding New Biases in Language Models with a Holistic Descriptor Dataset From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models HELM: Bias HELM: Language (Twitter AAE) HELM: Toxicity MASSIVE MT-bench - Benchmarking (with human and model scoring) The Self-Perception and Political Biases of ChatGPT Towards Measuring the Representation of Subjective Global Opinions in Language Models
Privacy Enhanced
HELM: Copyright llmprivacy mimir
Safe
Big-bench: Convince Me Big-bench: Truthfulness HELM: Reiteration, Wedging Mark My Words MLCommons The WMDP Benchmark

Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic (continued).

Secure and Resilient
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations detect-pretrain-code In-The-Wild Jailbreak Prompts on LLMs JailbreakingLLMs llmprivacy mimir TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
Valid and Reliable
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity Big-bench: Context Free Question Answering Big-bench: Contextual question answering, Reading comprehension, Question generation Big-bench: Morphology, Grammar, Syntax Big-bench: Out-of-Distribution Big-bench: Paraphrase Big-bench: Sufficient information Big-bench: Summarization DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations Eval Gauntlet: Reading comprehension Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming Eval Gauntlet: Language Understanding Eval Gauntlet: World Knowledge Evaluation Harness: BLiMP Evaluation Harness: CoQA, ARC Evaluation Harness: GLUE Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA Evaluation Harness: MuTual Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness FLASK: Readability, Conciseness, Insightfulness HELM: Knowledge HELM: Language HELM: Text classification HELM: Question answering HELM: Reasoning HELM: Robustness to contrast sets HELM: Summarization Hugging Face: Fill-mask, Text generation - Benchmarking Hugging Face: Question answering Hugging Face: Summarization Hugging Face: Text classification, Token classification, Zero-shot classification MT-bench

C.2: Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk

Table 4: Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk. Note that all available evals maybe applied to Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration risks.

CBRN Information
Big-bench: Convince Me (specific task) - Benchmarking
Big-bench: Self-Awareness - Benchmarking
Big-bench: Truthfulness - Benchmarking
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation - Attack
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations - Benchmarking
detect-pretrain-code - Attack/Benchmarking
HELM: Narrative Reiteration, Narrative Wedging - Benchmarking (with human scoring)
HELM: Question answering, Summarization - Benchmarking
In-The-Wild Jailbreak Prompts on LLMs - Attack
JailbreakingLLMs - Attack
LLM Privacy - Attack
MIMIR - Benchmarking
Mark My Words - Benchmark
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs - Attack
Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration
An Evaluation on Large Language Model Outputs: Discourse and Memorization: Benchmarking (with human scoring, see Appendix B)
BELEBELE - Benchmarking
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World - Benchmarking
Big-bench: Analytic entailment (specific task), Formal fallacies and syllogisms with negation (specific task), Entailed polarity (specific task) - Benchmarking
Big-bench: Context Free Question Answering - Benchmarking
Big-bench: Contextual question answering, Reading comprehension, Question generation - Benchmarking
Big-bench: Convince Me (specific task) - Benchmarking
Big-bench: Creativity - Benchmarking
Big-bench: Emotional understanding, Intent recognition, Humor - Benchmarking
Big-bench: Low-resource language, Non-English, Translation - Benchmarking
Big-bench: Morphology, Grammar, Syntax - Benchmarking
Big-bench: Out-of-Distribution Robustness - Benchmarking
Big-bench: Paraphrase - Benchmarking
Big-bench: Self-Awareness - Benchmarking
Big-bench: Social bias, Racial bias, Gender bias, Religious bias - Benchmarking
Big-bench: Sufficient information - Benchmarking
Big-bench: Summarization - Benchmarking
Big-bench: Toxicity - Benchmarking
Big-bench: Truthfulness - Benchmarking
BloombergGPT - Benchmarking
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation - Attack
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations - Benchmarking
DecodingTrust: Fairness - Benchmarking
DecodingTrust: Machine Ethics - Benchmarking (Machine Ethics Evaluation)
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations - Benchmarking
DecodingTrust: Stereotype Bias - Benchmarking
DecodingTrust: Toxicity - Benchmarking
detect-pretrain-code - Attack/Benchmarking
Eval Gauntlet Reading comprehension - Benchmarking
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming - Benchmarking
Eval Gauntlet: Language Understanding - Benchmarking
Eval Gauntlet: World Knowledge - Benchmarking

Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk. Note that all available evals maybe applied to Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration risks (continued).

Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration (continued)

Evaluation Harness: BLiMP - Benchmarking
 Evaluation Harness: C-Eval (Chinese evaluation suite), MGSM, Translation - Benchmarking
 Evaluation Harness: CoQA, ARC - Benchmarking
 Evaluation Harness: CrowS-Pairs - Benchmarking
 Evaluation Harness: ETHICS - Benchmarking
 Evaluation Harness: GLUE - Benchmarking
 Evaluation Harness: HellaSwag, OpenBookQA - General commonsense knowledge, TruthfulQA - Factuality of knowledge - Benchmarking
 Evaluation Harness: MuTual - Benchmarking
 Evaluation Harness: PIQA, PROST - Physical reasoning, MC-TACO - Temporal reasoning, MathQA - Mathematical reasoning, LogiQA - Logical reasoning, SAT Analogy Questions - Similarity of semantic relations, DROP, MuTual – Multi-step reasoning - Benchmarking
 FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness - Benchmarking (with human and model scoring)
 FLASK: Readability, Conciseness, Insightfulness - Benchmarking (with human and model scoring)
 Finding New Biases in Language Models with a Holistic Descriptor Dataset - Benchmarking
 From Pretraining Data to Language Models to Downstream Tasks:
 Tracking the Trails of Political Biases Leading to Unfair NLP Models - Benchmarking
 HELM: Bias - Benchmarking
 HELM: Knowledge - Benchmarking
 HELM: Language (Twitter AAE) - Benchmarking
 HELM: Language - Benchmarking
 HELM: Memorization and copyright - Benchmarking
 HELM: Miscellaneous text classification - Benchmarking
 HELM: Narrative Reiteration, Narrative Wedging - Benchmarking (with human scoring)
 HELM: Question answering - Benchmarking
 HELM: Question answering, Summarization - Benchmarking
 HELM: Reasoning - Benchmarking
 HELM: Robustness to contrast sets - Benchmarking
 HELM: Summarization - Benchmarking
 HELM: Toxicity - Benchmarking
 HELM: Toxicity detection - Benchmarking
 Hugging Face: Conversational - Benchmarking
 Hugging Face: Fill-mask, Text generation - Benchmarking
 Hugging Face: Question answering - Benchmarking
 Hugging Face: Summarization - Benchmarking
 Hugging Face: Text classification, Token classification, Zero-shot classification - Benchmarking
 In-The-Wild Jailbreak Prompts on LLMs - Attack
 JailbreakingLLMs - Attack
 LLM Privacy - Attack
 LegalBench - Benchmarking (with algorithmic and human scoring)
 MASSIVE - Benchmarking
 MIMIR - Benchmarking
 MT-bench - Benchmarking (with human and model scoring)
 Mark My Words - Benchmark
 Putting GPT-3’s Creativity to the (Alternative Uses) Test - Benchmarking (with human scoring)
 TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs - Attack
 The Self-Perception and Political Biases of ChatGPT - Benchmarking
 Towards Measuring the Representation of Subjective Global Opinions in Language Models - Benchmarking

Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk. Note that all available evals maybe applied to Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration risks (continued).

Environmental

An Evaluation on Large Language Model Outputs: Discourse and Memorization - Benchmarking (with human scoring, see Appendix B)
 BELEBELE - Benchmarking
 Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World - Benchmarking
 Big-bench: Analytic entailment (specific task), Formal fallacies and syllogisms with negation (specific task), Entailed polarity (specific task) - Benchmarking
 Big-bench: Context Free Question Answering - Benchmarking
 Big-bench: Contextual question answering, Reading comprehension, Question generation - Benchmarking
 Big-bench: Creativity - Benchmarking
 Big-bench: Emotional understanding, Intent recognition, Humor - Benchmarking
 Big-bench: Low-resource language, Non-English, Translation - Benchmarking
 Big-bench: Morphology, Grammar, Syntax - Benchmarking
 Big-bench: Out-of-Distribution Robustness - Benchmarking
 Big-bench: Paraphrase - Benchmarking
 Big-bench: Social bias, Racial bias, Gender bias, Religious bias - Benchmarking
 Big-bench: Sufficient information - Benchmarking
 Big-bench: Summarization - Benchmarking
 Big-bench: Toxicity - Benchmarking
 BloombergGPT - Benchmarking
 DecodingTrust: Fairness - Benchmarking
 DecodingTrust: Machine Ethics - Benchmarking (Machine Ethics Evaluation)
 DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations - Benchmarking
 DecodingTrust: Stereotype Bias - Benchmarking
 DecodingTrust: Toxicity - Benchmarking
 Eval Gauntlet Reading comprehension - Benchmarking
 Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming - Benchmarking
 Eval Gauntlet: Language Understanding - Benchmarking
 Eval Gauntlet: World Knowledge - Benchmarking
 Evaluation Harness: BLiMP - Benchmarking
 Evaluation Harness: C-Eval (Chinese evaluation suite), MGSM, Translation - Benchmarking
 Evaluation Harness: CoQA, ARC - Benchmarking
 Evaluation Harness: CrowS-Pairs - Benchmarking
 Evaluation Harness: ETHICS - Benchmarking
 Evaluation Harness: GLUE - Benchmarking
 Evaluation Harness: HellaSwag, OpenBookQA - General commonsense knowledge, TruthfulQA - Factuality of knowledge - Benchmarking
 Evaluation Harness: MuTual - Benchmarking
 Evaluation Harness: PIQA, PROST - Physical reasoning, MC-TACO - Temporal reasoning, MathQA - Mathematical reasoning, LogiQA - Logical reasoning, SAT Analogy Questions - Similarity of semantic relations, DROP, MuTual – Multi-step reasoning - Benchmarking
 Evaluation Harness: ToxiGen - Benchmarking
 FLASK: Background Knowledge - Benchmarking (with human and model scoring)
 FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness - Benchmarking (with human and model scoring)
 FLASK: Metacognition - Benchmarking (with human and model scoring)
 FLASK: Readability, Conciseness, Insightfulness - Benchmarking (with human and model scoring)
 Finding New Biases in Language Models with a Holistic Descriptor Dataset - Benchmarking
 From Pretraining Data to Language Models to Downstream Tasks:
 Tracking the Trails of Political Biases Leading to Unfair NLP Models - Benchmarking
 HELM: Bias - Benchmarking
 HELM: Knowledge - Benchmarking
 HELM: Language (Twitter AAE) - Benchmarking
 HELM: Language - Benchmarking
 HELM: Memorization and copyright - Benchmarking
 HELM: Miscellaneous text classification - Benchmarking
 HELM: Question answering - Benchmarking
 HELM: Reasoning - Benchmarking
 HELM: Robustness to contrast sets - Benchmarking
 HELM: Summarization - Benchmarking
 HELM: Toxicity - Benchmarking
 HELM: Toxicity detection - Benchmarking

Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk. Note that all available evals maybe applied to Confabulation, Dangerous or Violent Recommendations, Data Privacy, Human-AI Configuration, Information Integrity, Information Security, Intellectual Property, Toxicity, Bias, and Homogenization, and Value Chain and Component Integration risks (continued).

Environmental (continued)
Hugging Face: Conversational - Benchmarking
Hugging Face: Fill-mask, Text generation - Benchmarking
Hugging Face: Question answering - Benchmarking
Hugging Face: Summarization - Benchmarking
Hugging Face: Text classification, Token classification, Zero-shot classification - Benchmarking
Hugging Face: Conversational - Benchmarking
Hugging Face: Fill-mask, Text generation - Benchmarking
Hugging Face: Question answering - Benchmarking
Hugging Face: Summarization - Benchmarking
Hugging Face: Text classification, Token classification, Zero-shot classification - Benchmarking
LegalBench - Benchmarking (with algorithmic and human scoring)
MASSIVE - Benchmarking
MT-bench - Benchmarking (with human and model scoring)
Putting GPT-3’s Creativity to the (Alternative Uses) Test - Benchmarking (with human scoring)
The Self-Perception and Political Biases of ChatGPT - Benchmarking
Towards Measuring the Representation of Subjective Global Opinions in Language Models - Benchmarking
Obscene, Degrading, and/or Abusive Content
BELEBELE - Benchmarking
Big-bench: Convince Me (specific task) - Benchmarking
Big-bench: Low-resource language, Non-English, Translation - Benchmarking
Big-bench: Self-Awareness - Benchmarking
Big-bench: Social bias, Racial bias, Gender bias, Religious bias - Benchmarking
Big-bench: Toxicity - Benchmarking
Big-bench: Truthfulness - Benchmarking
DecodingTrust: Fairness - Benchmarking
DecodingTrust: Stereotype Bias - Benchmarking
DecodingTrust: Toxicity - Benchmarking
Eval Gauntlet: Language Understanding - Benchmarking
Evaluation Harness: C-Eval (Chinese evaluation suite), MGSM, Translation - Benchmarking
Evaluation Harness: CrowS-Pairs - Benchmarking
Evaluation Harness: ToxiGen - Benchmarking
Finding New Biases in Language Models with a Holistic Descriptor Dataset - Benchmarking
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models - Benchmarking
HELM: Bias - Benchmarking
HELM: Language (Twitter AAE) - Benchmarking
HELM: Memorization and copyright - Benchmarking
HELM: Narrative Reiteration, Narrative Wedging - Benchmarking (with human scoring)
HELM: Question answering, Summarization - Benchmarking
HELM: Toxicity - Benchmarking
HELM: Toxicity detection - Benchmarking
LLM Privacy - Attack
MASSIVE - Benchmarking
MIMIR - Benchmarking
Mark My Words - Benchmark
The Self-Perception and Political Biases of ChatGPT - Benchmarking
Towards Measuring the Representation of Subjective Global Opinions in Language Models - Benchmarking

Appendix D: List of Common Adversarial Prompting Strategies

D.1: Common Adversarial Prompting Strategies by Trustworthy Characteristic

D.2: Common Adversarial Prompting Strategies by Generative AI Risk

Appendix E: Common Risk Controls for Generative AI

E.1: Common Risk Controls for Generative AI by Trustworthy Characteristic

E.2: Common Risk Controls for Generative AI by Generative AI Risk

Appendix F: Example Low-risk Generative AI Measurement and Management Plan

6.3 F.1: Example Low-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.4 F.2: Example Low-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix G: Example Medium-risk Generative AI Measurement and Management Plan

6.5 G.1: Example Medium-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.6 G.2: Example Medium-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix H: Example High-risk Generative AI Measurement and Management Plan

6.7 H.1: Example High-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.8 H.2: Example High-risk Generative AI Measurement and Management Plan by Generative AI Risk